CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Approximating the Length of Chinese Postman Tours

Nathalie Bostel
Philippe Castagliola
Pierre Dejax
André Langevin

July 2013

CIRRELT-2013-42

Bureaux de Montréal :
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

UNIVERSITÉ LAVAL · UQÀM · HEC MONTRÉAL · ÉCOLE POLYTECHNIQUE MONTRÉAL · Université de Montréal

Approximating the Length of Chinese Postman Tours

Nathalie Bostel[1], Philippe Castagliola[1], Pierre Dejax[2], André Langevin[3,*]

[1] LUNAM Université, Université de Nantes et IRCCyN UMR CNRS 6597, Nantes, France

[2] École des Mines de Nantes, La Chantrerie, 4, rue Alfred Kastler, B.P. 20722, F-44307 Nantes, Cedex 3, France

[3] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal, C.P. 6079, succursale Centre-ville, Montréal, Canada H3C 3A7

Abstract. This article develops simple and easy-to-use approximation formulae for the length of a Chinese Postman Problem (CPP) optimal tour on directed and undirected strongly connected planar graphs as a function of the number of nodes and the number of arcs for graphs whose nodes are randomly distributed on a unit square area. These approximations, obtained from a multi-linear regression analysis, allow to easily forecast the length of a CPP optimal tour for various practical combinations of number of arcs and nodes ranging, from 10 to 300 nodes and 15 to 900 arcs.

Keywords. Vehicle routing, logistics, statistics, transport.

Acknowledgements. The authors wish to acknowledge the support of the Commission Permanente de Coopération Franco-Québécois (CPCFQ) and the Natural Sciences and Engineering Research Council of Canada (NSERC). They also thank Mr. Thomas Pleyber for his contribution to develop the graph generator.

---

* Corresponding author: Andre.Langevin@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2013

Introduction

Two important solution approaches for logistics and transportation problems are based respectively on mathematical programming and continuous approximations. The former approach relies on modeling and development of numerical methods requiring detailed data collection, whereas the latter relies on concise summaries of data and the development of analytic models. (Geoffrion, 1976) advocates the use of simplified analytic models to gain insights into numerical mathematical programming models. In a similar spirit, (Hall, 1986) illustrates applications of discrete and continuous approximations, and notes that continuous approximations are useful to develop models that are easy for humans to interpret and comprehend. Both authors agree that continuous models should supplement mathematical programming models but should not replace them. The article by (Newell, 1973) is considered as the seed of the continuous approximation approach to distribution problems. The reader is referred to the monograph of (Daganzo, 2005) for a pedagogical presentation of the continuous approximation methods and to (Langevin et al., 1996) for an overview of continuous approximation models that have been developed for freight distribution problems.

Distance approximations for multi-stop vehicle routes play a key role in many continuous models of transport and distribution of freight and persons. In the seminal work of (Beardwood et al., 1959) an asymptotic equation for the optimal length of a traveling salesman tour visiting N points distributed randomly in a region of area A is developed. From this equation, a formula is derived to approximate the length L of the optimal tour: $L = k\sqrt{AN}$ or equivalently $L = k\sqrt{N}$ on a unit square area. The constant k depends on the metric. The approximation is considered good for $N > 15$. Distance approximations for multiple stop peddling routes in distribution of goods are also developed by (Christofides and Eilon, 1969), (Eilon et al., 1971), (Daganzo, 1984b) and (Daganzo, 1984a).

In the domain of services, many routing problems correspond to arc routing

problems, e.g., waste management, snow disposal, meter collecting, and postman tours. No approximation formula for the length of arc routing tours has yet been developed. The basic problem in arc routing is the Chinese Postman Problem (CPP), introduced by (Guan, 1962). The CPP consists in finding the shortest closed tour that traverses all the edges and/or arcs of a graph at least once. For completely directed or completely undirected graphs there exists a polynomial algorithm. For mixed graphs, the problem is NP-hard.

The objective of this article is to develop approximation formulae for the length of a CPP optimal tour in the directed and the undirected cases for planar graphs. These formulae are functions of the number of nodes and arcs for graphs whose nodes are randomly distributed on a unit square area. The fitness of the formulae is evaluated statistically.

The article is organized as follows. The second section presents the methodology used. The third section describes how the graphs are generated. The fourth section presents the results and the statistical analyses and the last section provides some concluding remarks and future research directions.

Methodology

Our methodology consists in generating a large set of graphs and, for each graph, to optimally solve the Chinese Postman Problem in order to obtain the length of the CPP tour. We look for an approximate formula to predict the length y of the tour from the number of nodes x_N and the number of arcs or edges x_A, using a regression type method. Finally a statistical analysis is conducted to assess the validity of the formula. In the next section we describe the graph generation process.

Graph generation

Considering that an important number of various graphs is needed to conduct the experimentations, a graph generator that can generate random graphs with

specific characteristics was built. Strongly connected planar graphs are required.

Graph generation procedure

We use the following procedure to generate the directed graphs. For each graph we know the number of nodes x_N and the number of arcs x_A to generate.

1. Randomly generate x_N nodes in a unit square.
2. Find a Hamiltonian circuit that connects all the nodes.
3. Add arcs until the required number is obtained.

The undirected graphs are obtained from the directed ones by replacing each arc by its equivalent edge. Two arcs in reverse direction (e.g., arcs (a, b) and (b, a)) are replaced by a single edge.

The program to randomly locate the nodes on a unit square area is written in C++ and uses some LEDA (`http://www.mpi-inf.mpg.de/LEDA/`) library subroutines. Then, to insure that the graph is strongly connected, we generate a Hamiltonian tour that connects all the nodes, using a TSP algorithm (see `http://mathsrv.ku-eichstaett.de/MGF/homes/grothmann/java/TSP/`). The Grothmann heuristics finds the shortest tour by local descent from random positions. This heuristics ensures that the graph of the Hamiltonian circuit is planar. Then it is necessary to add arcs to the graph previously generated to reach the required number of arcs. We use the following algorithm to add arcs to the graph:

1. Randomly pick a node in the graph;
2. For this node, identify the list of possible nodes to be linked with, to constitute a valid arc, within a maximum distance. A valid arc must be a non-existing one; it must not intersect with another existing arc (to maintain the planarity of the graph); the degree of the destination node must be less than a given threshold;
3. If no such node exists, increase the maximum distance allowed and go to step 2.

Iterations are done until no more arc can be added or when the required number of arcs is obtained.

Graph characteristics

A database of 3600 directed graphs and 2700 undirected graphs has been created using the previous procedure. Because of the graph generation procedure that combines two reverse arcs in one edge to get the undirected graphs, it was not possible to generate undirected graphs with a ratio "number of edges over number of nodes" equal to 3. The generated graphs have the following input characteristics:

- number of nodes x_N, between 10 and 300, incremented by step of 10,

  number of arcs x_A equal to the number of nodes x_N multiplied by a coefficient that is respectively 1.5, 2, 2.5, and 3 for the directed graphs and 1.5, 2, 2,5 for the undirected graphs.
- 30 instances are generated for each pair of node and arc numbers,
- directed and undirected instances,
- arc length are computed using the euclidian metric.

For each generated graph, the following characteristics are recorded:

- the number of nodes (x_N),
- the number of arcs (x_A),
- the mean degree of the nodes,
- the standard deviation of the node degree,
- the length of the network (sum of all the arc lengths),
- the length of the optimal CPP tour.

Solving the CPP

For each generated graph, the optimal length of the CPP tour is determined using the classical optimal approach proposed by (Guan, 1962). In the directed case, it consists in solving a transportation problem by a standard linear programming solver (Xpress-MP) and, in the undirected case, a matching problem has to be solved using (Edmonds, 1965)'s algorithm.

Results

We performed a statistical study of the generated 3600 directed graphs and 2700 undirected graphs with the Scicoslab (`http://www.scicoslab.org/`) software, using a multi-linear regression analysis. We present the estimated coefficients corresponding to the model of equation (1) below for the directed and undirected cases in Tables 1 and 2 respectively.

To visualize the fitness of the proposed approximations, we present a graphical representation of the initial data compared with the curve obtained with the estimated parameters, in Figure 1 for the directed case and in Figure 2 for the undirected case. The abcissa values correspond to the number of nodes whereas the ordinate values correspond to the optimal length of the CPP. Four sets of points can be identified for the directed case, corresponding to the four ratios (1.5, 2, 2.5, and 3) considered between the number of arcs and nodes. Three sets of points are identified for the undirected case, corresponding to the three ratios (1.5, 2, and 2.5). To better analyse the fitness between the approximations and the data sets, we present the results for each set of observations independently. Each graph now represents the length of the CPP tour as a function of the number of nodes. Figure 1 for the directed case and Figure 2 for the undirected case show that the estimated models are quite relevant. Nevertheless, the dispersion of data is more important in the directed case than in the undirected case. This is conforming to our intuition that for the directed graphs the direction of the arcs has an impact on the variability of the total length.

Choice of the model

The goal of this section is to present some estimated models for the length y of the CPP tour as simple functions of the number of nodes x_N and the number of arcs x_A, for both the directed and undirected cases. The parameters of these models are estimated from the results obtained on the 3600 and 2700 generated graphs of the benchmark. Since we do not have any *a priori* knowledge about what the real model is, the key idea is to start with a very general and flexible model like the following one:

$$y(x_N, x_A) = a_0 + a_N z_N + a_A z_A + a_{NA} z_N z_A + \varepsilon \tag{1}$$

where a_0, a_N, a_A, and a_{NA} are 4 unknown parameters, ε is an error term, and z_N and z_A are new variables obtained using (Box and Cox, 1964) transformations of parameters c_N and c_A, respectively, i.e.

$$\begin{aligned} z_N &= \frac{x_N^{c_N} - 1}{c_N} \\ z_A &= \frac{x_A^{c_A} - 1}{c_A} \end{aligned}$$

The advantage of this simple model is that, depending on the values of c_N and c_A, it can exhibit polynomial and/or power type characteristics. Let nb be the number of graph instances (i.e., $nb = 3600$ for the directed case and $nb = 2700$ for the undirected case). The algorithm used for estimating the 6 parameters is the following:

STEP 1. Let $\mathbf{y}$ be the following column vector:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{nb} \end{pmatrix}$$

where y_i, $i = 1, 2, \ldots, nb$, is the length of the CPP tour corresponding to the ith experiment in the benchmark (directed or undirected cases) and let $\bar{y} = \frac{1}{nb} \sum_{i=1}^{nb} y_i$ be the average value of the y_i, $i = 1, 2, \ldots, nb$.

STEP 2. Set parameters c_N and c_A to some initial values like, for instance, $c_N = 1$ and $c_A = 1$ (i.e. the initial model is a simple linear + interaction model).

STEP 3. Compute $z_{N,i} = \frac{x_{N,i}^{c_N}-1}{c_N}$ and $z_{A,i} = \frac{x_{A,i}^{c_A}-1}{c_A}$, for $i = 1, 2, \ldots, nb$, where $x_{N,i}$ and $x_{A,i}$ are the number of nodes and the number of arcs, respectively, corresponding to the ith experiment in the benchmark (directed or undirected cases).

STEP 4. Compute matrix $\mathbf{X}$

$$\mathbf{X} = \begin{pmatrix} 1 & z_{N,1} & z_{A,1} & z_{N,1}z_{A,1} \\ 1 & z_{N,2} & z_{A,2} & z_{N,2}z_{A,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{N,nb} & z_{A,nb} & z_{N,nb}z_{A,nb} \end{pmatrix}$$

STEP 5. Estimate the column vector of parameters $\mathbf{a} = (a_0, a_N, a_A, a_{NA})^T$ using $\mathbf{a} = \mathbf{C}\mathbf{X}^T\mathbf{y}$ where $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$.

STEP 6. Compute the column vector $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{nb})^T = \mathbf{X}\mathbf{a}$ of estimated lengths of the CPP tour.

STEP 7. Compute the coefficient of determination $R^2 = 1 - \frac{SSE}{SST}$ where $SSE = \sum_{i=1}^{nb}(y_i - \hat{y}_i)^2$ and $SST = \sum_{i=1}^{nb}(y_i - \bar{y})^2$ are the Sum Squares of Error and Total, respectively. It is worth noting that the coefficient of determination $R^2 \in [0, 1]$ is a statistic that gives some information about the goodness of fit of a model and, in our case, how well model (1) approximates the lengths of the CPP tour. The larger R^2 the better the fit.

STEP 8. Change the values of c_N and c_A using a non-linear optimization algorithm ((Nelder and Mead, 1965) in our case) and loop to STEP 3 until R^2 reaches its maximum.

When it exits, this algorithm provides estimates for the 6 parameters a_0, a_N, c_N, a_A, c_A, a_{NA} maximizing the coefficient of determination R^2.

Model corresponding to (1)

	Estimates	CI 95%	p-value
a_0	-281.767	$[-352.298, -211.236]$	$< 10^{-6}$
a_N	3.9192×10^{-4}	$[4.6597 \times 10^{-5}, 7.3724 \times 10^{-4}]$	0.02613
a_A	194.941	$[191.717, 198.164]$	$< 10^{-6}$
a_{NA}	-3.3038×10^{-6}	$[-1.2206 \times 10^{-5}, 5.5982 \times 10^{-6}]$	0.4669
c_N	2.5879	-	-
c_A	0.4531	-	-

Model corresponding to (1) with $a_{NA} = 0$

	Estimates	CI 95%	p-value
a_0	-305.632	$[-373.934, -237.329]$	$< 10^{-6}$
a_N	9.9268×10^{-4}	$[6.3276 \times 10^{-4}, 1.3526 \times 10^{-3}]$	$< 10^{-6}$
a_A	199.77	$[196.396, 203.144]$	$< 10^{-6}$
c_N	2.3554	-	-
c_A	0.4478	-	-

Table 1: Directed graph case: estimated values for parameters a_0, a_N, c_N, a_A, c_A, a_{NA}

Directed case

Concerning the directed graph case, the estimated values for parameters a_0, a_N, c_N, a_A, c_A, a_{NA} are presented in Table 1 (top). 95% Confidence Intervals (CI) and p-values are also provided for a_0, a_N, a_A and a_{NA}. The optimal value for R^2 is 0.9249. Based on the estimated values of Table 1 (top), the corresponding estimated model is

$$y_{D1}(x_N, x_A) \simeq -281.767 + 3.9192 \times 10^{-4} z_N + 194.941 z_A - 3.3038 \times 10^{-6} z_N z_A$$

with

$$z_N = \frac{x_N^{2.5879} - 1}{2.5879} \quad \text{and} \quad z_A = \frac{x_A^{0.4531} - 1}{0.4531}$$

Since the p-value of a_{NA} in Table 1 is very large ($0.4669 >> 0.05$), this indicates that pararameter $a_{NA} = -3.3038 \times 10^{-6}$ has no influence on the model and thus the term in $z_N z_A$ can be omitted. The estimated values for parameters a_0, a_N, c_N, a_A, c_A assuming $a_{NA} = 0$ are also in Table 1 (bottom) with their corresponding 95% confidence intervals and p-values. The optimal value for R^2 is 0.9248 (i.e. almost unchanged compared to the model with $a_{NA} \neq 0$). Based on these new values, the reduced estimated model is

$$y_{D2}(x_N, x_A) \simeq -305.632 + 9.9268 \times 10^{-4} \left(\frac{x_N^{2.3554} - 1}{2.3554} \right) + 199.77 \left(\frac{x_A^{0.4478} - 1}{0.4478} \right)$$

In Figure 1 we have plotted the benchmark data corresponding to the directed graph case ($\circ$) for $x_N = 10, 20, \ldots, 300$ and for (a) $x_A = 1.5 \times x_N$, (b) $x_A = 2 \times x_N$, (c) $x_A = 2.5 \times x_N$ and (d) $x_A = 3 \times x_N$. We have also plotted the estimated model $y_{D2}(x_N, x_A)$ in plain line and the 95% confidence interval for the data in dotted lines. As it can be noted, the model $y_{D2}(x_N, x_A)$ fits the benchmark data very well, no matter the combination of (x_N, x_A).

Undirected case

Concerning the undirected graph case, the estimated values for parameters a_0, a_N, c_N, a_A, c_A, a_{NA} are presented in Table 2 (top) with their corresponding 95% confidence intervals and p-values. The optimal value for R^2 is 0.9947. Based on

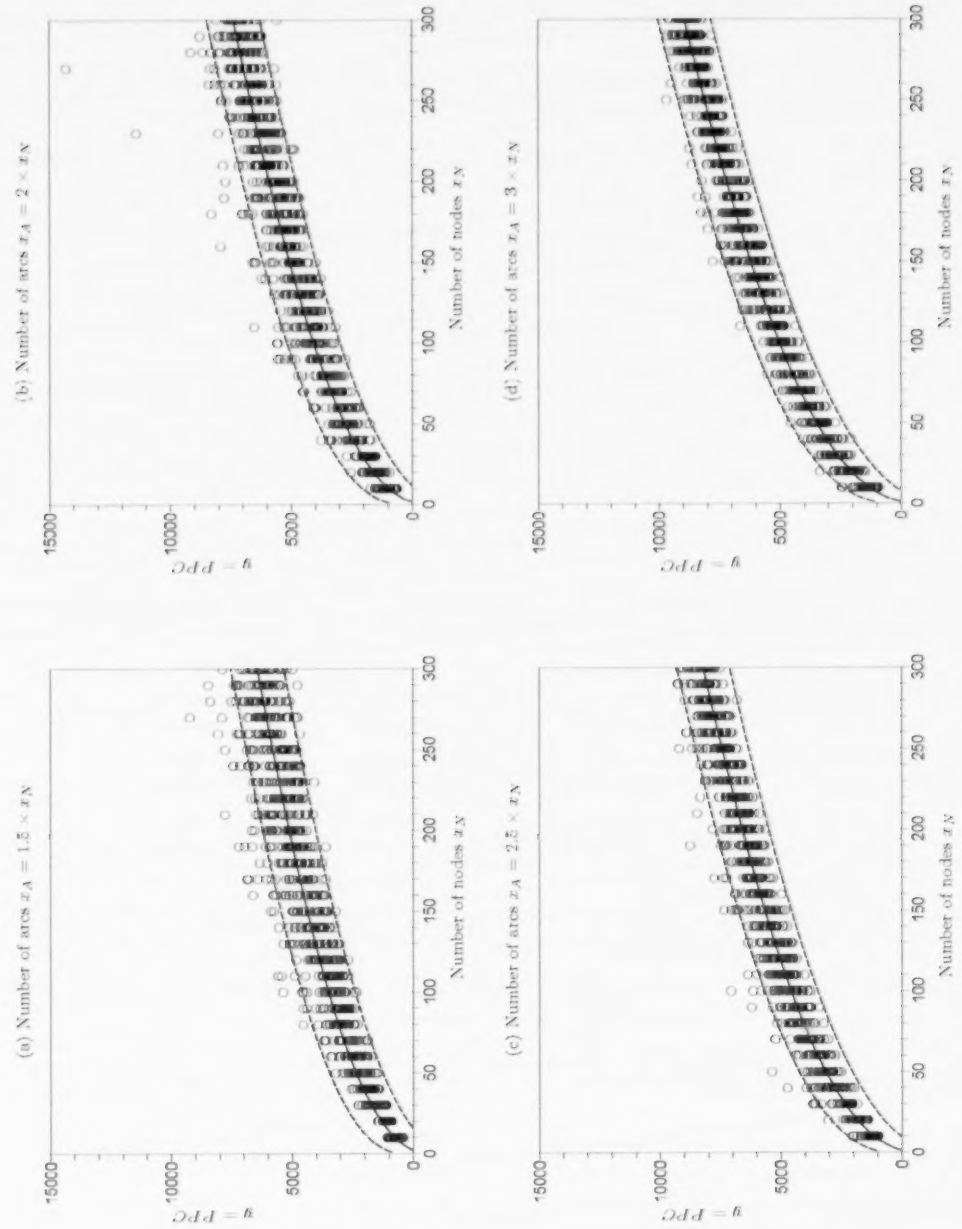

Figure 1: Comparison between the estimated model and the benchmark data for the directed graphs

Model corresponding to (1)

	Estimates	CI 95%	p-value
a_0	46.918	$[-166.304, 260.141]$	0.6662
a_N	91.414	$[-124.291, 307.118]$	0.4061
a_A	110.049	$[109.355, 110.744]$	$< 10^{-6}$
a_{NA}	-107.273	$[-107.952, -106.594]$	$< 10^{-6}$
c_N	-0.9748	-	-
c_A	1.4908	-	-

Model corresponding to (1) with $c_N = -1$ and $c_A = 1.5$

	Estimates	CI 95%	p-value
a_0	-356.154	$[-581.817, -130.491]$	0.00199
a_N	512.77	$[279.267, 746.273]$	0.00002
a_A	116.562	$[115.825, 117.299]$	$< 10^{-6}$
a_{NA}	-116.547	$[-117.286, -115.808]$	$< 10^{-6}$
c_N	-1	-	-
c_A	1.5	-	-

Table 2: Undirected graph case: estimated values for parameters a_0, a_N, c_N, a_A, c_A, a_{NA}

the estimated values of Table 2 (top), the corresponding estimated model is

$$y_{U1}(x_N, x_A) \simeq 46.918 + 91.414 z_N + 110.049 z_A - 107.273 z_N z_A$$

with

$$z_N = -\frac{x_N^{0.9748} - 1}{0.9748} \quad \text{and} \quad z_A = \frac{x_A^{1.4908} - 1}{1.4908}$$

In this case, it is worth to note that the parameters c_N and c_A are close to -1 and 1.5. For this reason, we have recomputed the parameters a_0, a_N, a_A and a_{NA} assuming $c_N = -1$ and $c_A = 1.5$. The results are shown in Table 2 (bottom) with their corresponding 95% confidence intervals and p-values. The optimal value for R^2 is 0.9946 (i.e. almost unchanged compared to the full model). Based on these new values, the simplifed estimated model is

$$y_{U2}(x_N, x_A) \simeq -356.154 + 512.77 z_N + 116.562 z_A - 116.547 z_N z_A$$

with

$$z_N = 1 - \frac{1}{x_N} \quad \text{and} \quad z_A = \frac{x_A^{1.5} - 1}{1.5}$$

In Figure 2 we have plotted the benchmark data corresponding to the undirected graph case ($\circ$) for $x_N = 10, 20, \ldots, 300$ and for (a) $x_A = 1.5 \times x_N$, (b) $x_A = 2 \times x_N$ and (c) $x_A = 2.5 \times x_N$. We have also plotted the estimated model $y_{U2}(x_N, x_A)$ in plain line and the 95% confidence interval for the data in dotted lines. As it can be noted, the model $y_{U2}(x_N, x_A)$ fits the benchmark data very well, no matter the combination of (x_N, x_A).

Remark : The formulae derived in this paper are relative to directed and undirected graph generated on a unit square. For the general case of graphs extended over a region area of surface A, the length of optimal tours provided by our formulae would have to be multiplicated by $\sqrt{A}$.

Conclusions

This article develops approximation formulae for the length of a Chinese Postman optimal tour on directed and undirected strongly connected planar graphs. The estimated length of the optimal tour is function of the number of nodes

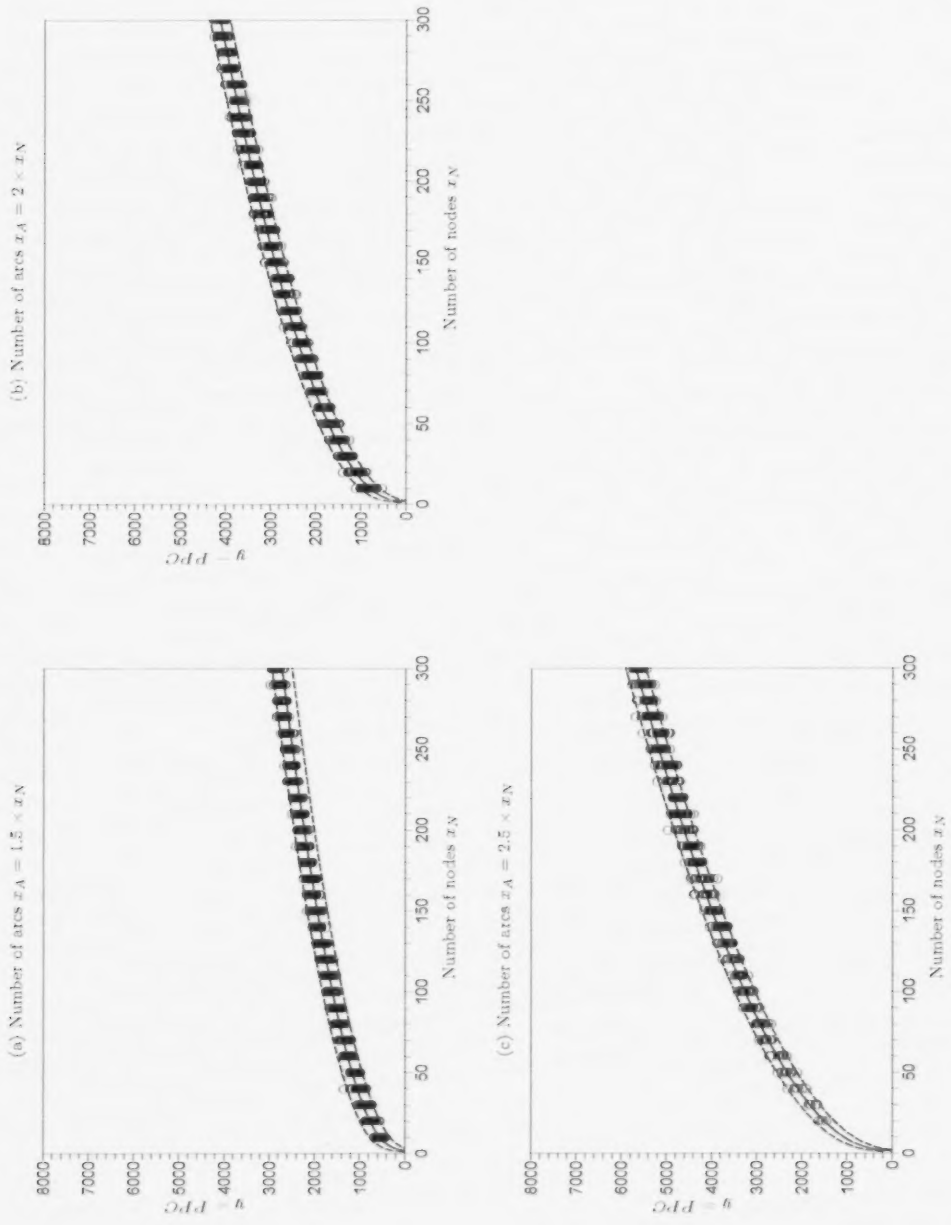

Figure 2: Comparison between the estimated model and the benchmark data for the undirected graphs

and the number of arcs for graphs whose nodes are randomly distributed on an unit square area. Using the actual optimal length of the Chinese Postman tour of 3600 directed graphs and 2700 undirected graphs, the coefficients of the formulae were estimated using a regression type method. A statistical analysis assessed the validity of the formulae which can be used to forecast the length of a CPP optimal tour for various practical combinations of number of arcs and nodes ranging from 10 to 300 nodes and 15 to 900 arcs.

Further research could be devoted to the adaptation of our methodology towards the development of approximation formulae for extended cases (non planar or not strongly connected graphs, or mixed directed / undirected graphs, non euclidean metrics). An interesting research avenue would be to apply our methodology to develop an approximation formula for the TSP on a network and compare it with the approximation formula of (Beardwood et al., 1959) on the plane. However this would require solving to optimality thousands of TSP. A challenging theoretical research direction could consider the determination of asymptotically exact formulae in a similar fashion to the (Beardwood et al., 1959) formula for the travelling salesman problem on the plane.

Acknowledgements

The authors wish to ackknowledge the support of CPCFQ (Commission Permanente de Coopération Franco-Québécois) and of the Natural Sciences and Engineering Research Council of Canada. They also thank Mr. Thomas Pleyber for his contribution to develop the graph generator.

References

J. Beardwood, J.H. Halton, and J.M. Hammersley. The Shortest Path through Many Points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

G.E.P. Box and D.R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B*, 26(2):211–252, 1964.

N. Christofides and S. Eilon. Expected Distances in Distribution Problems. *Operational Research Quarterly*, 20(4):437–443, 1969.

C.F. Daganzo. *Logistics Systems Analysis*, volume 36 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, 4th edition edition, 2005.

C.F. Daganzo. The Length of Tours in Zones of Different Shapes. *Transportation Research B: Methodological*, 18(2):135–145, 1984a.

C.F. Daganzo. The Distance Travelled to Visit n Points with a Maximum of c Stops per Vehicle: an Analytic Model and an Application. *Transportation Science*, 18(4):331–350, 1984b.

J. Edmonds. Paths, Trees, and Flowers. *Canadian Journal of Mathematics*, 17: 449–467, 1965.

S. Eilon, C.D.T. Watson-Gandy, and N. Christofides. *Distribution Management: Mathematical Modelling and Practical Analysis*, volume 36. Hafner, New York, 1971.

A.M. Geoffrion. The Purpose of Mathematical Programming is Insight not Numbers. *Interfaces*, 7(1):81–92, 1976.

M. Guan. Graphic Programming using Odd and Even Points. *Chinese Mathematics*, 1:273–277, 1962.

R.W. Hall. Discrete Models/Continuous Models. *Omega - The International Journal of Management Science*, 14(3):213–220, 1986.

A. Langevin, P. Mbaraga, and J.F. Campbell. Continuous Approximation Models in Freight Distribution: An Overview. *Transportation Research Part B: Methodological*, 30(3):163–188, 1996.

J. Nelder and R. Mead. A Simplex Method for Function Minimization. *Computer Journal*, 7(4):308–313, 1965.

G.F. Newell. Scheduling, Location, Transportation and Continuum Mechanics: Some Simple Approximations to Optimization Problems. *SIAM Journal on Applied Mathematics*, 25(3):346–360, 1973.